Davide Pavone

Data analysis of a transportation company

Helsinki Metropolia University of Applied Sciences

Bachelor of Engineering

Media Engineering

Data analysis of a transportation company

Date Helsinki, 6 April 2016

Author	Davide Pavone
Title	Data analysis of a transportation company
Number of Pages	37 pages
Date	Helsinki, 6 April 2016
Degree	Bachelor of Engineering
Degree Programme	Media Engineering
Instructor	Aarne Klemetti, Researching Lecturer

The purpose of this thesis project was to investigate the database of a Finnish taxi company in order to provide them a summary of their performance achieved in a specific three-month period. The scope of the research is to understand the data structure, reorganize the database, analyse sets of data, show results, limits and eventually deficiencies in order to give useful information of the performances of the taxis.

The study started by getting familiar with the information the database contains. The information concerns specific rides for customers with special needs. Trips are organised by a company working for the city of Helsinki and the information is collected by an application. It was necessary to adopt a model to use as a structure for the whole project and CRISP-DM model was chosen. For this study, RStudio open source environment with different packages was chosen as workspace. After introducing different concepts regarding data and illustrating models for data mining, the research focuses on analysing a segment of the database.

The results yielded with the support of different and interactive graphs from different perspective can help the company to interpret the profitability of working with these special rides provided by the city.

Keywords	Data, database, data analysis, big data, RStudio, CRISP-DM.

Contents

Terms and Abbreviations

API	Application Programming Interface
CRAN	Comprehensive R Archive Network
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSV	Comma-Separated Value
EOS DIS	Earth Observing System Data Information System
ETL	Extract-Transform-Load
GNU	GNU is Not Unix
ID	Identification
IDE	Integrated development Environment
IT	Information Technology
KDD	Knowledge Discover Database
NASA	National Aeronautics and Space Administration
NCR	National Cash Register
NISO	National Information Standards Organization
SEMMA	Sample Explore Modify Model Assess
SPSS	Statistical Package for Social Science (IBM)
XLS	Microsoft Excel spreadsheet file

1 Introduction

During the coldest week of this winter, one day I noticed that the windshield washer fluid of my car got frozen so I went to visit a shop specialised in car spares parts and supplies. Surprisingly, the department where I was supposed to find different options was almost empty. The few choices available were products suitable for temperatures until -18°C, yet that week it was much colder. The situation was similar in the department for car batteries. Probably, the shop was not able to estimate enough in advance the upcoming need. With a comparison of weather forecast and products needed even a car supplies shop can improve the business. This simple example is just to tell how important it is nowadays to analyse data.

For the purpose of this research a decision was made to apply the power of RStudio environment, by using different packages, to a database provided by a taxi company in Helsinki. The goals for this project are:

- Show which of the taxis is performing best and explain the reason;
- Predict where rides are requested;
- Discover if some day of the week there are more rides;
- Present if there is specific period of the day when more taxis are required;
- Produce a variety of graphic depictions summarising the data, taking in consideration different perspectives;
- To understand if it is useful to use RStudio for this kind of project.

The scope of this thesis is to illustrate how to solve the most common issues when starting working with a new dataset, show a possible first approach how to read the data, define and use a work environment in order to provide results that can help to improve the performance of a company.

2 Project

In 2014 a taxi company in Helsinki owning several taxis stipulated a contract with a company organising services for the city. The object is providing taxi service for customers that have special needs, often with physical or mental challenges. The amount of the trips that the customer can have differs from individual to individual. Some of the customers also have a small fee to pay for each trip.

These special rides are implemented with the help of an application. The customer needs to book a car some time in advance and an operator in the office is organising the trip with the help of a software while a taxi driver by using a special application, can see the taxi request. The taxi company provides different cars for this service and there are three different kinds of taxis. There is a car driving normal rides and special rides, a car driving only these special rides for a few hours per day and a car having an 8,5-hour of shift and they are allowed to drive only these special rides. The software used by the operator stores all the information in a database. A copy of this database is provided to the taxi company each month. The database contains mostly sensible data such as driver ID, customer name, addresses, different fees, but also kilometres driven, time of the trip and much more. For this research a small copy of the database containing over thirty-four thousands of rows, in “xls” format, was provided covering time from June 2014 until January 2015. The sensitive data such as driver ID, customer name or fee of the ride was omitted in order to guarantee privacy. Under inspection of the thesis, only the nine taxis driving these special rides were taken into consideration.

3 Methods

In this chapter after introducing meanings and definitions of terms related with big data and clarify different types of data I will illustrate the environment and language used to develop this project and reasons that have influenced my choice to use RStudio and R language. Also, the CRISP-DM model, which will be illustrated later, was used as a structure to develop a data mining project for this research.

3.1 Big data and little data

Nowadays, big data has become a hot topic and has received attention in the public opinion. However, little data is important too. Big data is not necessarily referring to high volume of data. A better quality of data and having the correct data is actually more relevant. However, very often the problem is that there are no data at all for several reasons such as that the data can not be accessed or the data are not significant or they are so corrupted that they can not be used.

In the 2013 the Oxford English Dictionary introduced the term big data defining it as: *"data of a very large size typically to extent that its manipulation and management present significant logistical challenges"*. Another way to interpretat the meaning of big data is referring to the value the data can represent, so big or little data refer to the importance of the result that data can reveal. [1, 3-10.]

The power law distribution, used as a function relation to illustrate the relationship between two quantities in statistics, called *the long tail* by Chris Anderson illustrated in the magazine Wired is another way to represent the use of data. This theory was introduced in 2004 to represent that a small number of commercial products, so called "hits", is very popular (the *Head* of the representation) whereas a large number of niche products have much less popularity (The *Long Tail*) as shown in figure 1.

The meaning of the power law distribution applied to the use of data is that a small number of users (corresponding to the product in the long tail representation) are dealing with large volumes of data (popularity), a large number of users are dealing with very little data and most of the users are somewhere in between. This description is

used to highlight that in the reality there is a restricted number of fields where a large number of data is involved. [2.]

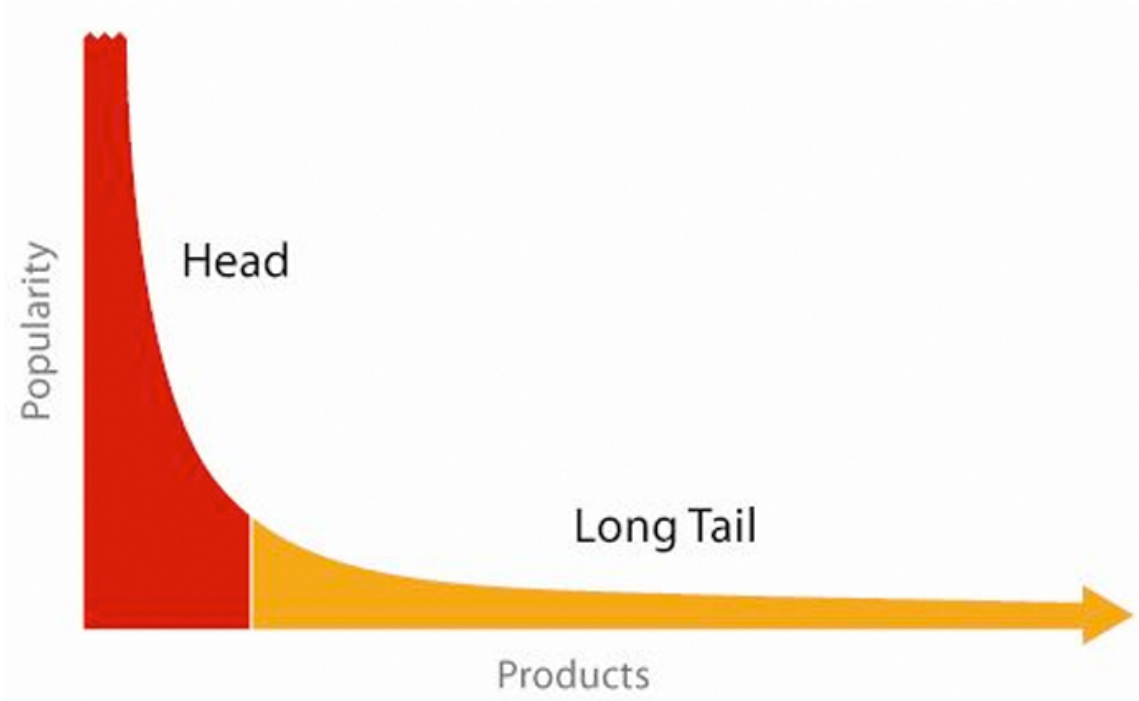

Figure 1. The power law distribution known as “the long tail”. Reprinted from Anderson(2004). [2].

3.1.1 What and when are data?

It is interesting to know that the word data is not a new word. According the Oxford English Dictionary it was already in the plural form in 1646 and it was used in theology and the first uses were in Latin. Later it was defined as *“the set of principles accepted as the basis of an argument”*. There are several interpretations and meanings of this word. A more relevant aspect than *“what are data?”* is actually *“when are data?”*, referring to a process that create data. Linked to this concept, Ray Harris and Peter Fox gave a definition where they say: *“ 'Data' includes, at a minimum, digital observation, [...] visualizations, and statistical data collected for administrative or commercial purposes. Data are generally viewed as input to the research process”*.[3.]

Continuing to explore other definitions in order to give a better picture of the meaning of data I found the studies of Paul Uhlir and Daniel Cohen defining that data can be generated by people or machines or communication between software and models [1,17-19].

Especially when working for research purposes, it could be practical to categorize types of data in groups. One way of doing this is by grouping by degrees of processing. Earth Observing System Data Information System (EOS DIS) by NASA groups their data in five groups of processing levels. This specific division is relevant for different needs required by the scientists working with that data. The different levels of processing are from the level with data cleaned, to the level at full resolution to an other level that includes processed data. These levels of processing can be adapted and personalised to any project concerning data analysis not necessarily related to large amount of data. [1, 21-22.]

3.1.2 Types of data

As already mentioned before when talking about big data, we do not necessarily mean the overall size. Another relevant aspect is the scope of the research that can influence type and size of the data and how they are collected. The process of data collection, in scientific fields for example, can involve sophisticated machines or specific software that are developed as a result of elaborate research. Also, an individual or a limited group of people can collect data for projects. The most extreme way is an artisanal data collection where information is manually compiled and collected by researchers.

There are types of information already suitable for research with no need of intervention, while others demand some kind of manipulation such as cleaning or other kind of adjustment.

Another distinction in diversity of data lies in the origin of data. Were the data collected directly or indirectly by the scientist or did it already exist? This distinction divides *sources* and *resources* data. Sources data are generated for a specific project by members of the team, sometimes they are also called new data. While resources means already existing data reiterated for a new research, these are also called old data. Nonetheless, a project based in resources data may be combined with new data, sources data. [1, 58-65.]

One of the first things when starting to work with a set of data is classification. This step is crucial for organizing the data and it helps to develop the project better. First of all, in classification we need to give names to the groups of data and later we can dispose

the categories in the most appropriate way. The outcome of a pertinent classification can help for a valid interpretation and investigation of the data. [1, 65-68.] These data are called metadata, literally *"data about data"* that means data that describes other data. The main role of metadata is to make the life of the researcher easier when looking for particular information about specific data. Sometimes the information included in the metadata is just to locate an object or document in the environment, for example, the code added in the cover of a book located in the library. In some other case metadata includes information that describes an object better. In the case of the book, for example, there can also be information about the author. Other examples of metadata are data which are created or modified or file size. These kinds of metadata are very useful for video or picture archives. [4.]

A study conducted by the NISO (National Information Standards Organization) classified metadata in three types: descriptive, structural and administral. *Descriptive metadata* illustrates data in order to better explore and classify them. *Structural metadata* reveals the relation between different objects, this is the case of how different pages of a book are sorted together in order to form the chapters. *Administrative metadata* contribute to give information to handle the data, for example, information about when a file was created and the format or who has the right to access it. Administrative metadata can be divided in other categories: *rights management metadata* that concerns the rights to the file and *preservation metadata* that includes those instructions to store the data. [5.]

Another relevant aspect in metadata is the *provenance*, term borrowed from the French that means the origin of something. The provenance reveals the history of that information. Especially in a project where data need to be preserved and archived this kind of metadata has a particular relevance. [6, 70.]

3.2 R language and RStudio as software environment

There are several software and work environments available for analysing data with different characteristics and based on different languages. RStudio was chosen for this project after other work environments were tested. RStudio may not the simplest work environment but already after a few steps, it gives a considerable impression how powerful it can be.

3.3 Overview of RStudio

RStudio is an integrated development environment (IDE) based in R, a programming language mostly used for statistics and data analysis. RStudo runs on desktop and it is available for Windows, Linux or Mac operating system. It can also run in an internet browser connected to the RStudio server. It is available in open source licence or commercial edition for organization.

The RStudio workspace is mainly divided in two resizable columns. Each column is divided in other two resizable parts. The upper window of the left column works mainly as source editor for writing R scripts and for displaying the datasets in use. The lower window of the left column is the console used to run the commands. The upper window of the right column is the workspace to view objects in the global environment. It is primarily used when loading a new dataset. This part of the workspace contains also a searchable command history. The lower part of the right column has several functions. It is mostly used for installing and updating packages, showing plots and it includes also an integrated R help. [2.] Figure 2 shows an overview of an RStudio workspace.

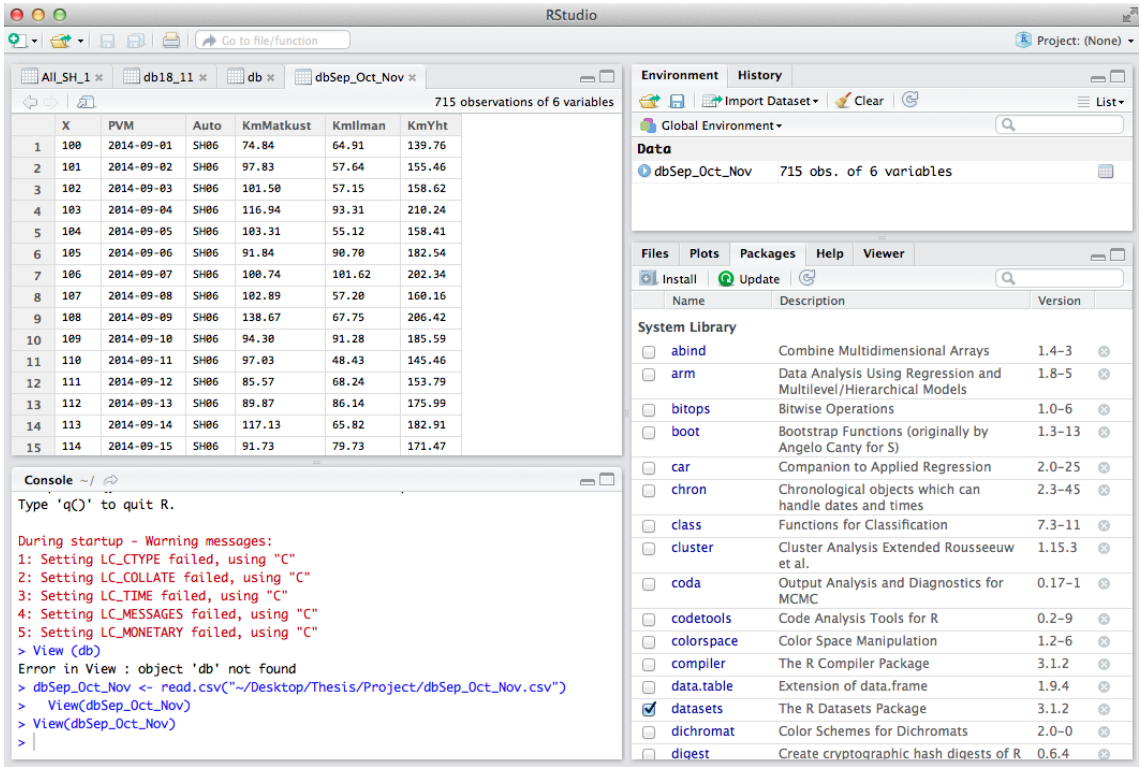

Figure 2. Overview of the RStudio workspace

3.4 R language

R is a programming language and environment developed for statistics and graphics purposes. It is a GNU project with General Public Licence. GNU (GNU is Not Unix) is an operating system distributed as free software where the users have the freedom to run, distribute and also improve the software. R is based on S language and most of the code written for S can run also in R. R has a large assortment of statistical and graphical tools that makes this language extremely powerful for data analysis. R can be extended via packages. R is considered a straightforward and effective language. This statement and the large variety of tools that R includes developed for data analysis and statistic aims were determinant for considering using as main work environment for studying this project. [7;8.]

3.5 Packages and tool for R

Especially for an open source such as R and RStudio it is important to find reliable and official information. There is a webpage called “r-bloggers.com” which provides users answers to any problem or support related to R environment. This web page can be like a headquarter for a company or at the same time like a dictionary for a translator where a new user but also R language programmers find and exchange information. R-bloggers.com is an aggregator of different contents written by experts of R language. [9.]

On this web page it is also possible to find basic guidances how to install and start using R and tutorials about the packages and other tools. Here are some of the most important packages that can be used for the basic research with RStudio:
Plyr is used for the basic splitting, applying and combining operation. **Datasets** is a package containing dataset for practising. **Devtools** helps to develop a new package and at the same time it allows users to download and install packages that are not on the CRAN (Comprehensive R Archive Network), the network collecting all the codes and documentation for R. [10.] **Plotly** released in 2015 is a powerful and interactive browser-based charting for visualising data. **Ggplot2** a very well documented and straithforward package developed for providing graphics representations.

3.6 Datamining

Before starting analysing the project it is time to say a few words to introduce the science that is meant for data analysis: data mining. According the Gartner IT glossary: "*Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical technique."* [11.] Basically with this process, scientists analyse data from a different point of view in order to try to provide relationships between different elements or set of data. Consequently, data can yield useful results that were ignored.

When working in a project involving data analysis, it could be very useful to follow a model or a process that acts as the main structure and guides the scientist to the whole procedure until a result is yielded. In the next sections three of the most used models are described: *KDD process*, *SEMMA model* and *CRISP-DM model*.

3.7 KDD process

The Knowledge Discover in Database (KDD) is a process used for extract data where data mining represents a fundamental section. This term was already introduced in 1989. This process comprises mainly five steps as illustrate in the figure below.
These five steps are:
Selection. One or more sub set of data (or target data) are created.
Preprocessing. Mainly cleaning the target data.
Transformation. Applying different methods to reduce the data.
Data Mining. Finding patterns that reveal new relevant aspects.
Interpretation/Evaluation. Interpretation of the pattern and evaluation of the process. [12;13.]
Figure 3 shows the different steps in the KDD process,

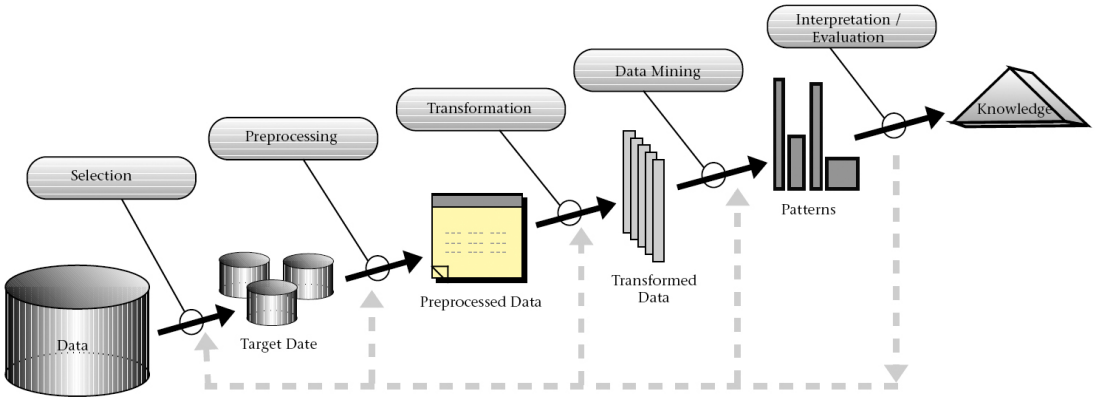

Figure 3. KDD process. Reprinted from Fayyad (1996) [12].

3.8 SEMMA model

Statistical Analysis Software Institute (SAS), one of the main software suites for data mining developed a very detailed own model, mainly designed for SAS Enterprise Miner software called Sample Explore Modify Model Assess model (SEMMA). As illustrated in the figure below, it basically included five steps:

***S**ample.* The subset extracted, from the original data, should be large enough to contain a sample with enough information but at the same time should be small enough to be managed.

***E**xplore.* This stage comprises understanding of the dataset. The scientist tries to anticipate the possible relationship between the attributes.

***M**odify.* In this phase the variables contained in the dataset are transformed based on the goal to be achieved.

***M**odel.* At this level several analytic methods and data mining tools are used to provide prediction and results.

Assess. Finally, the results provided are evaluated. Eventually some modification and more modeling is required to achieve more accurate results. [13; 14.]

Figure 4 shows the different phases of the SEMMA model.

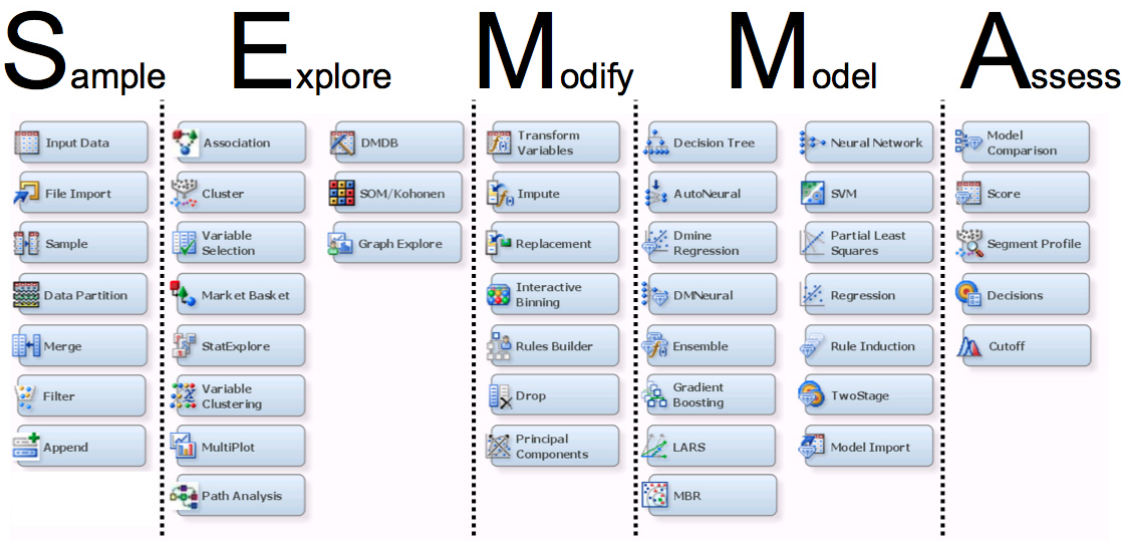

Figure 4. SEMMA Model. Reprinted from SAS (2015) [14].

3.9 CRISP-DM model

In 1996, scientists from DaimlerChrysler, SPSS and NCR developed a data mining process model that can be applied to different field of research. The model, called CRISP-DM (**Cr**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining) illustrates the life cycle of a project based in data mining. The indent of this model is to guide researchers to develop a project in a liner way and create a standard that helps to interpret the result achieved. The CRISP-DM consists of six phases. The relationships between these phases, as illustrates below, can be bidirectional.

The six phases are:

1. *Business and research understanding*. This phase concerns defining the goal of the project and delineating the strategy to carry out to obtain the goal.
2. *Data understanding.* This phase involves collecting the data, understanding and starting to explore the data. This stage is important also to inspect the quality of the data.
3. *Data preparation.* In most of the case the data acquired need to be elaborate. In this phase the raw data is processing by cleaning, transforming and all the other requiring operations, to provide a dataset for the following phase. The time consumed in this phase it depends very much on the quality of the data in relationship with the goal of the project.

4. *Modelling.* In this phase different techniques for modelling are applied. A some point of this step is very common going back to the preview phase and eventually is required a new dataset.
5. *Evaluation.* In this phase a model for the project is produced. It is also time for determining if the results achieved meet the goals of the project.
6. *Deployment.* This phase usually provide a report or require some more data mining in other department. Not necessary this phase means the end of the project, but most important is that the customer at this stage understands the actions required. [15.]

Figure 5 shows the phases of the CRISP-DM model.

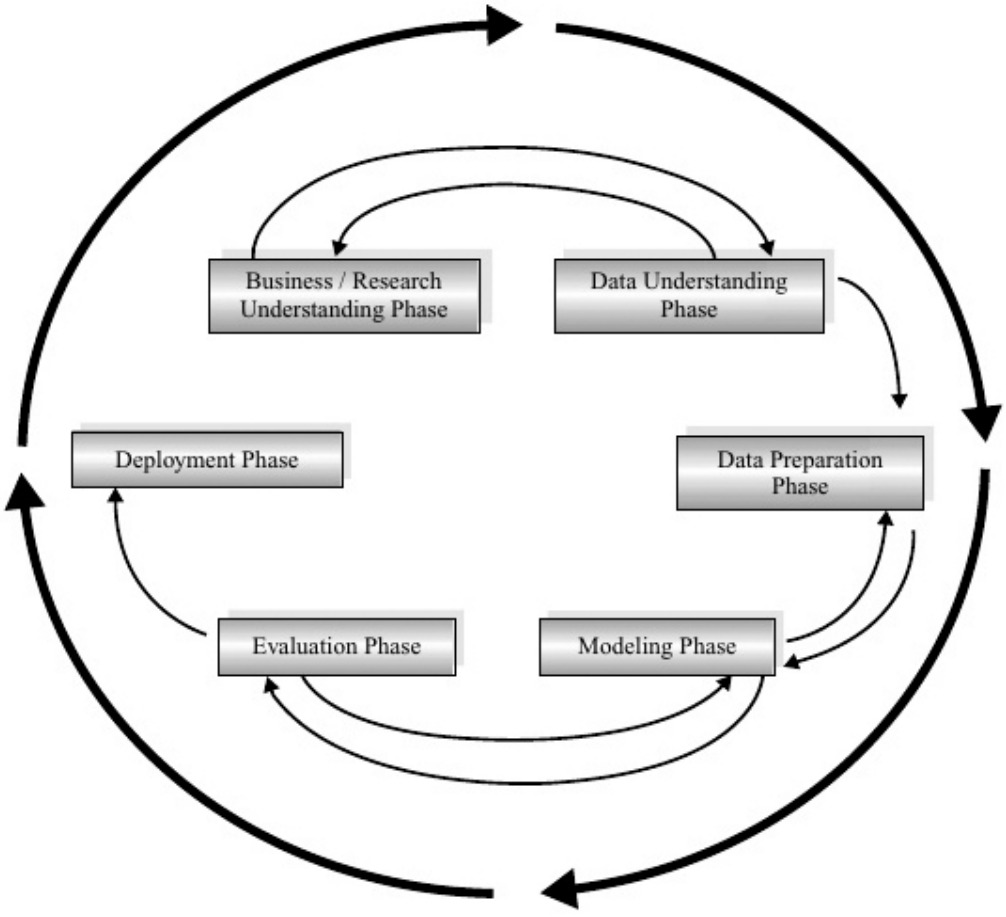

Figure 5. CRISP-DM model. Reprented from Larose (2004) [15].

4 Analysing and data mining

In this section I apply the CRISP-DM model to the project developed for this thesis. The main idea is to illustrate the fundamental steps of the analysis of the project and to explain the strategy used to achieve this goal. The main idea is to apply the theory already illustrated in the section 3.9 to the database provided by the taxi company. For better reading the evaluation and deployment, the phases are presented in separated chapters.

4.1 Extract-Transform-Load (ETL) process

Usually when talking about **E**xtract-**T**ransform-**L**oad (ETL) we are applying the process from the source system in order to store the information to the data warehouse. In this case we can borrow the same terminology for data mining by applying this process to the original database considering as source system, in order to get the data model.

The extract stage, consists on thus operations to carry out for making the data accessible. Usually includes cleaning steps where attributes such as date or name are changed in an appropriate format or null values are cutting out. The transform stage can include sorting, calculating and aggregating different attributes to get new values for better reading or for summarizing and reduce the size of the dataframe. Usually the purpose of load step is to provide a database or more for the modelling process. [16.]

4.2 Business and research understanding

It was not necessary to create a strategy for collecting the data used in this study as the data already existed. However, it was necessary to find a strategy how to use the data to provide results that could yield interesting unknown facts about the performance of the taxis. So, it is important to understand the data contained in the database in order to recognise what we can achieve with it. In this case, we can say that the research phase was subordinate to the *data understanding* phase.

After becoming familiar with the data we can make firmer the goals of the project that were already mentioned in the introduction. However, we can summarise that mainly we would like to understand if a particular taxi is performing better than another one

and what the possible reasons for it may be. The other, perhaps more interesting, question that we would like to answer concerns the predictability of the coming rides, whether it is possible to predict if a certain ride is coming with regularity. The results obtained are illustrated with graphs and a report.

4.3 Data understanding

When starting to work on a project where the subject to analyse a dataset, the first issue is, of course, to get the data. Since a dataset often contains sensitive information, privacy is of the upmost importance. In the case of this project, sensitive data such as the names of customer, the ID numbers of taxi drivers and fees of rides have been omitted in order to respect the privacy of the people.

For the purpose of this research we have used resourced data. When working with old data there are a few considerations to be taken in to account. First of all, it may take time to understand the meaning of the information contained in the data. Not necessarily all the data are required for the research, and more than likely a substantial part of the data will be disregarded during the cleaning process, in the data preparation phase. The goal of a project may be subordinated and adapted to the existing data assigned to the project. On the other hand, a project based on sourced data will, most probably, require much more time to acquire the desired data.

One of the most common formats for saving data, especially used for small datasets, is a spreadsheet format such as Excel. When receiving such files, the first step is to convert them in to a tab-separated text file such as CSV (comma-separated value), which is the most common text format for storing data; each line represents a record and each field is separated by a comma and some of the value can be enclosed in double quotes. The main reason this format is so popular is because it can be read by text and spreadsheet software and also by more specific software for statistic and data analysis [17].
Figure 6 and 7 show how it looks like a "csv" file.

```
"1","01/01/2015","SH06",49.24,16.23,65.47
"222","01/01/2015","SH07",95.66,84.28,179.95
"443","01/01/2015","SH08",42.37,40.89,83.24
"664","01/01/2015","SH09",60.09,45.55,105.65
"885","01/01/2015","SH21",69.77,63.36,133.13
```

Figure 6 Example of CSV file.

1	01/01/2015	SH06	49.24	16.23	65.47
222	01/01/2015	SH07	95.66	84.28	179.95
443	01/01/2015	SH08	42.37	40.89	83.24
664	01/01/2015	SH09	60.09	45.55	105.65
885	01/01/2015	SH21	69.77	63.36	133.13

Figure 7. CSV file opened with spreadsheets software.

4.4 Data Preparation

When for the first time we have a look at a dataset, we tend to assume that the data included look cleaned, probably because we are too enthusiastic to start analysing the content of the database. In most of the cases, when we start to understand the data structure of a dataset, there are a number of operations that might be required in order to get a much more readable dataset.

Usually, a large amount of the data contained in a database is not processed. As mentioned earlier, in a case where the data used are resources data, such as the data inspected for this research, the data provided are going to sustain a substantial drop in size before going under the process of modelling. Examples of most common data that are disregarded during this phase are as follows: redundant fields, missing values, data in a format not valid for data mining and not consistent values [15].

4.4.1 Inappropriate fields name

At the beginning of the project, after a quick look at the data set, one of the first inconsistencies noticed was the way used to assign a name in the different columns. Titles assigned in the columns gave a clear description of the data contained (figure 8). However, the length of the names were often too long and included spaces between the words. These two features generated difficulties and more appropriates names were required:

F
Reitin arvioitu päättymisaika (kellonaika)
08:59:00 AM
09:55:00 AM
10:44:00 AM

Figure 8. Example of original named assigned to a column of the dataset.

This kind of operation of changing objects name, part of the cleaning process, is easy to hold with R without using any other software by typing the following command:

```
names(db)[6]<-"KelloAlk"
```

Where “db” is the name of the dataset, “6” is the number of the column that has to be modified and “KelloAlk” is the assigned new name.

PVM	KelloAlk	KelloArv	KelloPäät
06/09/2014	08:40:00	09:15:00	08:54:00
06/09/2014	09:15:00	09:23:00	09:19:00
06/09/2014	09:30:00	09:45:00	09:35:00

Figure 9. Example of a new name assigned to a column.

4.4.2 Date format

As explained in the CRISP-DM model, sometimes more operations concerning the cleaning process are required after the model process. This is the case in regard to the date type format present in our dataset; apparently in the original format it seems to be working, and only at the modelling process step has an error been detected where the format “mm/dd/yyyy” is not allowed by the R system. The type of date format allowed is “mm-dd-yyyy”. The date format can be converted by the following function:

```
db$PVM = strptime (x=as.character(db$PVM), format= "%m/%d/%Y")
```

Where “db” is the name of the dataset that we are using and “PVM” is the name of the column with the format that has to be changed. The result is illustrated in the figure 8 and 9 below:

PVM
01/01/2015
01/01/2015

Figure 10. Date format not allowed

PVM
2015-01-01
2015-01-01

Figure 11. Date format allowed

4.4.3 Data cleaning and file exporting

In order to achieve the goals prefixed for this project only a part of the original dataset was required. In particular, only objects with the attribute starting with “SH” in the field “Auto” were considered valid. This specification shaped the original dataset from over 34,000 rows to a smaller dataset of over 21,000 rows.

An other typical operation required during the process of data preparation is the cutting out of irrelevant columns from the original dataset. This can be obtained with the command:

```
db2=db1[,c('Auto','PVM','KelloAlk','KelloArv','KelloPäät','KmMat
                    kust','KmIlman','KmYht')]
```

Here “db2” is the name of the new dataset that includes the only desired columns, and in this case “db1” is the dataset that included the original columns. In quotation marks, included in the brackets, is the name of the only columns required. In this way, the dataset quickly drops in size and become easier to read and process.

After these basic cleaning steps at the preparation data phase we are ready to deliver a first version of a much more readable dataset where the names of the columns were redesigned, the date format was changed to the correct form and undesired columns and rows were dropped out.
At this stage we can export from R a copy of the dataset created with the following command:

```
write.csv(db, "db.csv")
```

Hence, a file called “db.csv” in “.csv” format is made from the dataset “db”. Also, figure 12 illustrates how the dataset delivered for the modelling phase looks.

Auto	PVM	KelloAlk	KelloArv	KelloPäät	KmMatkus	KmIlman	KmYht	Reitti
SH06	2014-06-09	08:40:00	09:15:00	08:54:00	14.27	17.56	31.83	[YLLÄSTUNTURINTIE,HELSINKI],[HAAVIKKO RAUTAPORTTI Metsäpuronpolku 1,Helsinki]
SH06	2014-06-09	09:15:00	09:23:00	09:19:00	0.49	10.04	10.52	[MYLLYPURON OSTOSKESKUS KIVIPARINTIE 2,HELSINKI],[Karistimentie,Helsinki]
SH06	2014-06-09	09:30:00	09:45:00	09:35:00	1.87	0.71	2.58	[Myllypurontie,Helsinki],[TALLINNANAUKIO 1 DANSKE BANK,HELSINKI]
SH06	2014-06-09	09:48:00	10:42:00	00:00:00	13.84	13.74	27.58	[KONTULAN OSTOSKESKUS PARKKIPAIKKA R-KIOSK KONTULANKAARI 1,Helsinki],[Kivipyyk
SH06	2014-06-09	10:50:00	11:22:00	00:00:00	10.70	12.03	22.73	[TALLINNANAUKIO 1 DANSKE BANK,HELSINKI],[PRISMA VIIKKI VIIKINTORI 3 TAKSIRUUTU,

Figure 12. Database ready for the modelling phase.

4.5 Data modelling

If data preparation is the phase where the researcher usually starts to put his or her fingers on the keyboard and starts to explore RStudio or another platform, in the data modelling phase curiosity grows and the scientist tries to find answers to their questions. In data analysis the most enticing thing is trying to reveal something new or find some relation between data that helps to predict a further stage. However, data analysis is not just finding relationships, but it helps to provide other important documents such as graphs and reports that summarize the overall business.

4.5.1 Example of data corrupted

In the case of the dataset used for this project, the question that seems to be the most interesting and that can reveal important facts concerns the possibility of finding a relationship between cars and journeys; or in other words we are interested if it is possible to find an association between cars that always drive the same journeys. Another question that we want to immediately explore concerns the possibility to predict where and when cars are required and if there are always specific journeys that are required with regularity at the same time and from the same starting address to the same ending address. These two aspects are both related to the field called “Reitti” that includes information about the starting and ending point of a journey. This field, as shown in figure 11, reveals also other information. The first thing that is possible to notice is that the field “reitti” includes both starting and ending points inside different square brackets. The number of the building is not always included, as in some case we have only the name of the street. Things get more complicated because sometimes inside the square brackets we have also other information such as the instructions of where the taxi has to stop, the names of the buildings or shops. In some case we have four or even six square brackets; these are cases of multiple customers at the same time, so we can have up to three customers coming or going to the same or different places.
An attribute divulging this kind of information becomes quite difficult to be explored. This example of an attribute that includes corrupted and missing data could be quite common in such databases as it is not designed for purposes such as this research. In the case of a project utilising sources data where the database structure is designed from scratch, and is already at the research phase where a strategy of how to carry the project and collect the data has to be developed, this information should be contained

in a more specific way. For example, by having separate attribute collecting starting and ending addresses that include the building number and a separate attribute that stores driving information or building names.

After these considerations, we decided against using the attribute “Reitti”. This means that we are not able to answer research questions that predict when and from where similar taxis journeys are received.

[LAPINLAHDENKATU 16 - MARIAN SAIRAALA,Helsinki],[Tyynelänkuja,Helsinki]
[Jumbo SOL (Vantaanportinkatu 3),Vantaa],[LANDBONPOLKU,HELSINKI]
[Juhana-herttuan Tie,Helsinki],[CITYMARKET ITÄKESKUS SHELLIN PUOLI KAUPPAKARTANONKATU 3,Helsinki]
[KEINUTIE,HELSINKI],[Kontukuja,Helsinki]
[HAAPASAARENTIE,HELSINKI],[Rajatorpantie,Vantaa]
[Alihaka,Espoo],[Tarkkampujankatu,Helsinki],[Tarkkampujankatu,Helsinki],[Metsänpojankuja,Espoo]

Figure 13. Detail of the field “reitti”.

4.5.2 Narrow the range

At this stage we decided to model the original database more and provide a narrower dataset. In particular it was considered necessary to select a specific time where there was not a particular event, such as a holiday, in between. For this reason the range estimated was between 1 September and 30 November. Moreover, kilometres driven by the taxis were seen as being important to focus on since they are related to the performances. Before selecting a range between dates, a narrower dataset was established where each row includes information for each car for each day and the sum of kilometres that each car drove for each journey. In this way we could get the total amount of kilometres driven by a car in a day and not simply each journey. In doing so, we got a very compressed dataset of 715 rows.

The operation applied, called aggregation, works only if the date format is in the original format: “dd/mm/yyyy”. This means that the following operation has to be done before converting the date format:

```
db= aggregate(. ~ PVM+Auto, data = db, FUN = sum)
```

With this expression we can se that the column “PVM” and “Auto” are combined together and the other columns concerning kilometres are summed. After this operation

the attribute “PVM” about date can be converted as was already illustrated earlier and we can range the dataset as desired:

```
db[which(db$PVM > "2014-08-31" & db$PVM < "2014-12-01"),]
```

With this operation the dataset provided information only from 1 September until 30 November. Some of the taxis are also driving on Saturdays and Sundays, so in order to have more consistent data we are excluding these days from the dataset with the following operation:

```
db[(!db$PVM == "2014-09-06") & (!db$PVM == "2014-09-07") &
(!db$PVM == "2014-09-13") & (!db$PVM == "2014-09-14") & (!db$PVM
== "2014-09-20") & (!db$PVM == "2014-09-21") & (!db$PVM ==
"2014-09-07") & (!db$PVM == "2014-09-27") & (!db$PVM == "2014-
09-28") & (!db$PVM == "2014-10-04") & (!db$PVM == "2014-10-05")
& (!db$PVM == "2014-10-11") & (!db$PVM == "2014-10-12") &
(!db$PVM == "2014-10-18") & (!db$PVM == "2014-10-19") & (!db$PVM
== "2014-10-25") & (!db$PVM == "2014-10-26") & (!db$PVM ==
"2014-11-01") & (!db$PVM == "2014-11-02") & (!db$PVM == "2014-
11-08") & (!db$PVM == "2014-11-09") & (!db$PVM == "2014-11-15")
& (!db$PVM == "2014-11-16") & (!db$PVM == "2014-11-22") &
(!db$PVM == "2014-11-23") & (!db$PVM == "2014-11-29") & (!db$PVM
== "2014-11-30"),]
```

In this way the new dataset created included less than 585 rows from the original over 21,000 rows. The dataset is illustrated in figure 14:

PVM	Auto	KmMatkust	KmIlman	KmYht
2014-09-01	SH06	74.84	64.91	139.76
2014-09-02	SH06	97.83	57.64	155.46
2014-09-03	SH06	101.50	57.15	158.62
2014-09-04	SH06	116.94	93.31	210.24
2014-09-05	SH06	103.31	55.12	158.41

Figure 14. Dataset summarising information after aggregating cars and date.

From the new dataset created we can see there are nine taxis driving every day from Monday to Friday. They have different shifts of 8 hours and a half between five o'clock

in morning and nine o'clock in the evening. In some rare case they get a journey outside of their shift. One more consideration is that the driver can have a break for up to one hour.

To illustrate how many shifts or days each car is driving we can provide the following table, by running this code:

```
table1=table(db$Auto)
```

We create a table named “table1” from the dataset called “db”, and we are counting how many rows concern the same car. The result shown in figure 15 reveals that each car in the dataset is presents 65 times. This proves that each car is driving the same amount of time in the desired range of time, and it means that the data are consistent:

SH06	65
SH07	65
SH08	65
SH09	65
SH21	65
SH26	65
SH27	65
SH28	65
SH29	65

Figure 15. Table showing the number of shifts per taxi of the period selected.

4.5.3 Applying Tapply function

At this stage, after further modelling the dataset, we are ready to count the number of total kilometres that each car has been driving. In the dataset in use we have three different kinds of attributes dealing with kilometres (figure 14). We have “KmMatkust” that are the kilometres that a taxi is driving with the customer. “KmIlman” are the kilometres that a taxi is driving from the place where it is situated at the moment it receives the journey until the place where the taxi picks up the customer. Then we have “KmYht” that is the total amount of kilometres of the two previous attributes. Since the driver is paid also for the kilometres that are driven before picking up a customer, we will take into consideration the last attribute “KmYht” or the total kilometres. Also, those

kilometres are not the total kilometres driven by a taxi in a day; in effect the taxi driver does not get any fee for those kilometres that he freely drives from the moment he has arrived at the destination desired by the customer until he receives the next ride.

In data analysis it is very common to summarise variables by using a function to get the mean or sum of the variables. The “Tapply()” function is very practical, and it is used to break up vectors by a factor [18]. In our case we need to summarise the kilometres driven by each taxi during the three months. The function “tapply()” is quite straightforward as is possible to see below:

```
TotalKM=tapply(db$KmYht, db$Auto, FUN=sum)
```

As already mentioned, we decided to use the total kilometres driven by taxis, “KmYht” and we are interested in applying the function “sum” in order to get all the kilometres driven and grouped together by car “$Auto”. The result is a matrix that has to be converted in a data frame before proceeding to further steps:

```
b=as.data.frame(TotalKM)
```

Following this, the matrix is converted in a data frame, after is ordered by increasing order and it looks as in figure 16:

```
table=c[order(c$TotalKM),]
```

As shown in figure 16, with these steps we are able to answer the first question presented in the goal of the project: Which is the best taxi performance in the range of the three months selected?

	Total_KM
SH27	9250.09
SH26	10224.24
SH08	10431.66
SH28	10579.14
SH21	10808.87
SH06	10931.20
SH29	10949.12
SH09	11040.30
SH07	11968.03

Figure 16.Total kilometres driven by each taxi.

Taxi	Starts	Ends
SH27	08:00	16:30
SH26	05:00	13:30
SH08	08:00	16:30
SH28	06:30	15:00
SH21	11:30	20:00
SH06	08:00	16:30
SH29	12:30	21:00
SH09	11:00	19:30
SH07	13:00	21:30

Figure 17. Taxis shifts

As already mentioned all the taxis drove 65 shifts in three months between September and November. After a quick look at figure 14 we can see that cars drove between 9,250 km and 11,968 km, with an average of 10.687 kilometres per car as provided with the following function:

```
mn=mean(c$TotalKM)
```

Before make any conclusions we should take in consideration the different shifts that the taxis were driving (figure 17), but we can already see that most of the cars drove 400 kilometres more or less than the average; except taxis SH 27 and SH 07 that, respectively, drove roughly 1400 kilometres less and more than the average.

After checking the list of the taxi shifts, we can see that the three cars driving less have a shift in the morning and the three taxis driving more kilometres have a shift in the afternoon. Also, by comparing performances and shifts taxi SH27, that was the one driving less, has the same shift of taxis SH08 and Sh06 that drove, respectively, 1,200 and 1,700 kilometres more. It may be interesting to investigate more about this substantial difference that could be affected by different factors. One of these factors could be the driver. Basically, if they were always the same drivers with the same taxis we can deduce that some drivers are slower or faster than others or that someone is having less or more breaks than other drivers. Unfortunately, information such as driver ID is omitted from the database provided for this research for privacy reasons so it is not possible to investigate further.

4.5.4 The boxplot

So far, we have only analysed the result by making some deductions by viewing the numbers. A common approach used in statistics that provides a graphic illustration is called boxplot. This standard method does not show numbers and instead displays the distribution of data by presenting five values: the minimum, the first quartile, the median, the third quartile and the maximum. Data are split into quartiles. Data between the smallest point below the box and the lowest part of the box determines the first quartile (Q1). The part between Q1 and the line inside the box that represents the median is the second quartile (Q2). The third quartile is between Q2 and the highest part of the box, and finally we have the fourth quartile that reaches the largest point outside the box. The box is delimited by Q1 and Q3 and is called inter-quartile range (IQR), and the lines expanding from the box until the minimum and maximum value are called whiskers. [19.]

In R Studio we can provide a boxplot with the following code:

```
boxplot(c$TotalKM, col=c("darkred"))
```

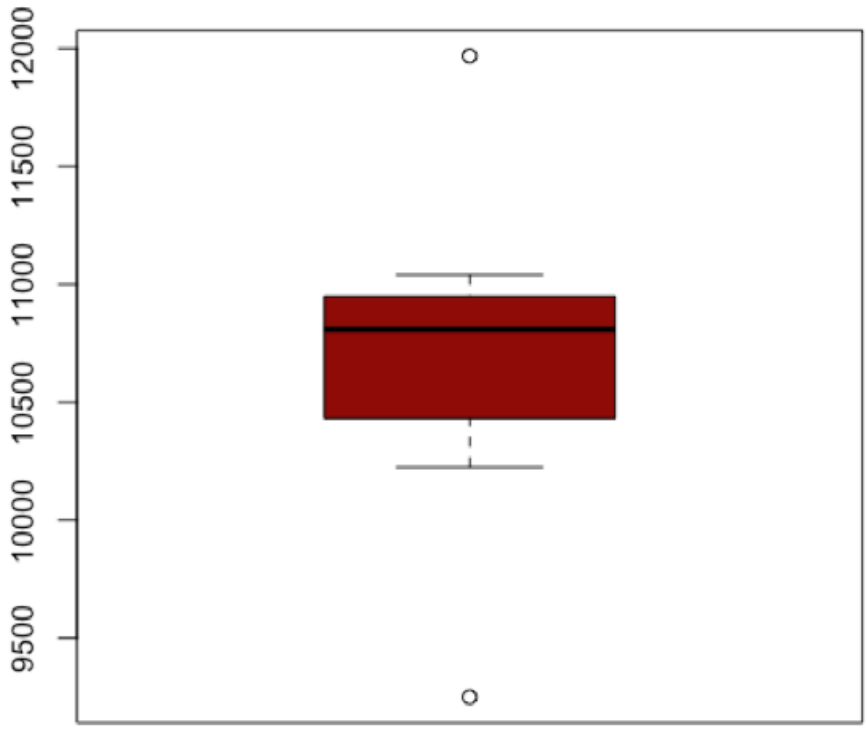

Figure 18. Plotbox about kilometres drove by the taxis.

The plot as shown in figure 18 gives an interesting representation of the result obtained. With the data displayed in such a way, we can spot the minimum and maximum value that in this case represents the taxis that have been driving less and more kilometres. The other two lines outside the box show the second lowest and highest value; in other words the second cars that have driven less and more kilometres. The scale on the left outside of the plot indicates the kilometres.
The plotbox is a common way to display results, especially when there is a lot of data to summarise.

4.5.5 Barchart with Plotly library

In regard to the result so far achieved, we do not have a lot of data to display; just the performances of nine cars. For this reason we wanted to utilize a more specific tool that can display all nine performances, and we decided to utilize a barchart by using a new library called Plotly, which was developed by the company Plotly based in Montreal. The API is available for different platforms in the fields of statistics such as R, MAT-

LAB, Python and Julia. The Plotly package is used for creating interacting web-based graphs and was released in November 2015. [20.]

After installing the Plotly library in RStudio by using another package called Devtool, developed particularly for helping in using new packages, we have to recall it in the console by running the following commands:

```
library (plotly)
p<- plot_ly()
p <-
plot_ly(x=c("SH06","SH07","SH08","SH09","SH21","SH26","SH27","SH
  28","SH29"),y=db(db$TotalKM), name="barchart",type = "bar")
```

As we can see from the command we are using the attribute “TotalKM” of the “db” dataframe. The result is an interacting bar chart as shown in figure 19.

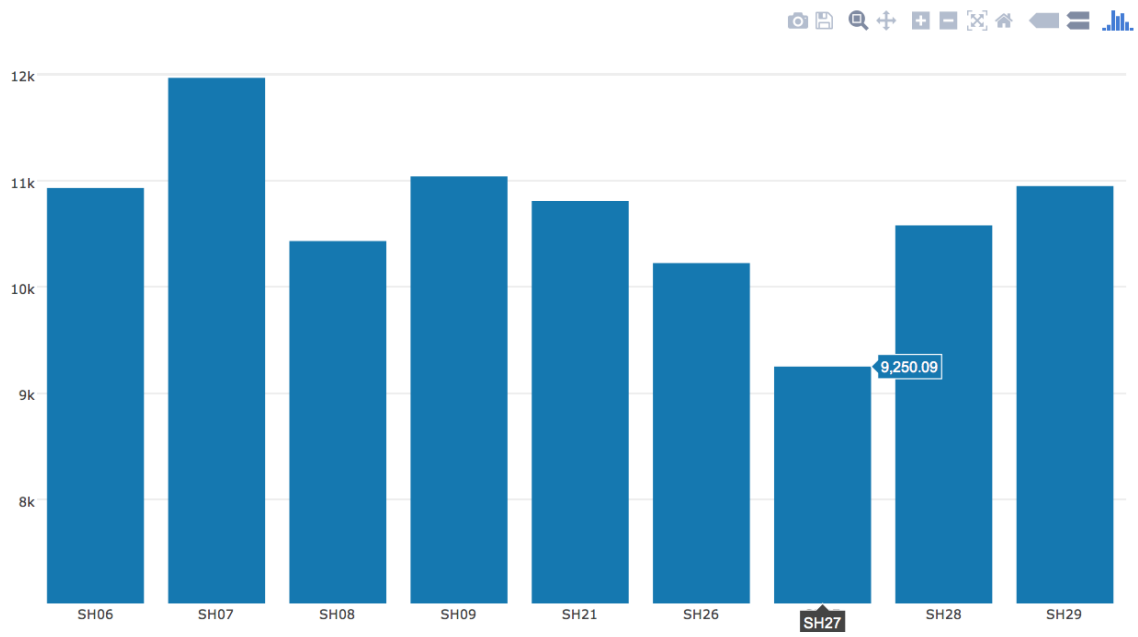

Figure 19. Bar chart showing total kilometres driven by each taxi.

On the y-axes of the barchart is the total number of kilometres expressed in “k” (thousand), and on the x-axes are the seven cars. The interaction of the barchart in this case is visible by moving the cursor of the mouse to show the total number of kilometres driven by the cars. In figure 19, the screenshot was taken with the cursor on top of the bar representing the car “SH27”. The barchart provided by Plotly has also other functions visible in the top corner of the right of the above picture: zooming, resizing, saving, editing and more.

4.5.6 Finding the number of rides driven by each taxi by using frequency

In statistics a common way to measure values is to use frequency. Basically, with frequency it is possible to measure how many times a certain type of event occurs. In our case we have several taxis driving several rides every day, so the number of the rides can be considered as frequency. Since the driver get also has a fixed compensation for each ride, this value can reveal important information to determinate the performances of each taxi.

This method essentially involves two attributes; the date and the name of the taxi. After using the frequency we obtain a database where we have only one row for each car for one day, and from this we get a new column with the number of how many rides in one day a taxi has been driven. This value represents the frequency. The commands used is as follows:

```
freq=data.frame ( table ( q$Auto, q$PVM ) )
```

With these commands we obtain the frequency in a new column by using the function “table”, and in this way a counter for how many rides are driven by the same taxi in one day is added. After this operation we can filter the data obtained, as already shown before, by selecting the same range of three months and excluding rows concerning Saturday and Sunday; this summarises the frequency and converting the table in data frame. Finally, we merge the result with the other data frame with the following command:

```
NewDataframe=merge(db,db1)
```

The “NewDataframe” is the name of the newly created dataframe and “merge” is a function that allows to combine columns from different dataframes: “db” and “db1” in this case. The result is illustrated in figure 20.

Auto	TotalKM	Rides
SH06	10931.20	868
SH07	11968.03	847
SH08	10431.66	892
SH09	11040.30	849
SH21	10808.87	814
SH26	10224.24	746
SH27	9250.09	833
SH28	10579.14	810
SH29	10949.12	822

Figure 20. Table with total amount of kilometres and number of rides.

By introducing the frequency we can have a more accurate analysis of the performances of the taxis. In general, we can say that there is a smaller gap between the taxis. For example, if we take into consideration the car" SH07" that drove about 1,500 kilometres more than "SH08", then the latter was driving 45 rides more. This gap tends to be even smaller if we take into consideration another important element that is the consumption of fuel.

4.5.7 Find out if there are more rides on some days of the week

One of the research questions was to try to find out if during the week there are days when taxis are driving more rides. The strategy to find an answer to this question starts from the original database through only selecting the period range already used (from September until November) and then excluding Saturday and Sunday. After we provide five different data frames, one for each day of the week, we noticed that there are thirteen dates for each day of the week in the range of the three months selected. The number of the rows included in each data frame represents the number of total rides. In this way we can see if there are more rides on some days of the week. For example, we can provide a data frame including all Monday's rides with these commands:

```
monday=bw1[(bw1$PVM=="2014-09-01") | (bw1$PVM=="2014-09-08") |
      (bw1$PVM=="2014-09-15") | (bw1$PVM=="2014-09-22") |
      (bw1$PVM=="2014-09-29") | (bw1$PVM=="2014-10-06") |
      (bw1$PVM=="2014-10-13") | (bw1$PVM=="2014-10-20") |
      (bw1$PVM=="2014-10-27") | (bw1$PVM=="2014-11-03") |
      (bw1$PVM=="2014-11-10") | (bw1$PVM=="2014-11-17") |
                  (bw1$PVM=="2014-11-24"),]
```

The operation used is similar to the one we used for excluding all Saturdays and Sundays. In this case we do not use the symbol “!” before the database “bw” (in this case). The main difference is that to add the conjunction “and” for selecting another date we have to use the symbol “|” whereas before for excluding we were using “&”.

After getting five different data frames, one for each day of the week, we use the following function to find out the number of rows that coincide with the number of total rides on that particular day of the week:

```
mo=NROW(monday)
```

For example, with this command we stored in “mo” the number of rows of all the rides drove on Monday. After collecting this information for each day of the week and converting the vector obtained in the dataframe, we merged all information as already show before in a data frame. The result is shown in figure 21.

Monday	Tuesday	Wednesday	Thursday	Friday
1501	1482	1510	1494	1494

Figure 21. Total number of rides for each day of the week during the period prefixed.

The result is very homogeneous. The total number of rides for each day of the week is very similar. After a quick look at the result, we can conclude that for each day of the week there is the same amount of customers in the three months selected. However, this conclusion is actually wrong. Most probably the reason for this result is that these special taxis are driving only rides provided by the city, which that means that they are not allowed to take any other customers, for example from the taxi station. So, the company that is organizing these rides for the city has to ensure a certain number of trips to the company providing the taxis. However, the result can be interpreted that rides are equally shared during the week in the range selected, and that the skill of the drivers does not affect much the overall performance of any of the nine taxis.

4.5.8 Analysing the kilometres driven every day by using ggplot2 package.

In order to gain a much clearer picture regarding the productivity of the different taxis, it is necessary to analyse individually the kilometres driven by each taxi day by day. We used data collected from nine cars in 65 days, and in order to be to visualise all the information at a glance we had to first find an interpretation that allowed us to summarise the information. The plot box already used, could be an important tool to illustrate the kilometres drove by each car every day. In this way we can get a plot box for each taxi summarising the kilometres drove each day.

In this case we used another package called “ggplot2”. This library, developed by Hadley Wickham, is considered one of the best graphic tools available for R. It can be used to show univariate and multivariate numerical and categorical values. The strength of this package is that values can be easily represented with elegant graphic, and values can be differentiated by colour, symbol, size and transparency [21]. This library is very well documented, and it is quite straightforward as it is possible to see from the function used:

```
library(ggplot2)
qplot(data = bwx, x=Auto, y=KmYht, color= Auto, geom="boxplot")+
  scale_color_manual(values=c("red", "lightblue", "green", "si-
 enna", "orange", "royalblue2", "purple", "yellow", "black")) +
  ggtitle("Representation of the kilometres driven every day by
                each taxi from September until
            Novembter")+labs(x="Taxi",y="Kilometres")
```

After loading the library we used the function “qplot”, and “bwx” is the name of the data frame. “Auto” is the column used that represents the x-axis. On the y-axis we will use the column “Yht” to represent the total kilometers driven each day by each car; this means that the data frame contained one row summarising the total kilometres driven per day by each taxi for a total amount of 585 rows. Then, we added also other function to personalise the colors “scale_color_manual”, the title of the graph “ggtitle” and the text added to the x and y-axis “labs”.

Figure 22. Box plot representing the kilometres driven each day by the taxis.

Each box plot in figure 22 contains information about the total kilometres driven every day by each taxi. The length of the box shows the range of total kilometres per day which the car has been driving the most. The end of the lines outside of the box is the whiskers represent the minimum or maximum kilometres that were driven in a day. In some cases there are also dots, lower or higher than the whiskers, which represent the total kilometres driven in a day. The reason for why these dots are so far from the box or the “whiskers” is because they show a much smaller or bigger value compared to the others values contained inside the box. The power of the box plot applied to these data is that at a glance it is possible to see the average kilometres driven by a taxi in a day. With this graphic depiction it is also possible to individuate the minimum and maximum kilometres driven in a day and to compare all the results of the taxis.

4.5.9 Scatter plot

Further using the package “ggplot2”, we decided to provide a more specific view of the kilometres driven by each taxi. In this case a graph that can be used is a scatter plot where each dot on the graph is a value. To produce a more readable representation the data frame is split into three different data frames; one per month. In this way, we

can get a graph where the y-axis shows the total kilometres driven by a taxi each day and the x-axis the time. The result is a graph easier to read where each car is represented by a different colour and each dot is linked by a line.

The package provides several functions that can be added to provide a personalized graph. Since on the graph nine taxis are represented, it was necessary to select manually the colours for the taxis; this function is called ”scale_color_manual”. The main function “qplot” recalls the name of the data frame to use, “sep” in this case, that includes only data regarding September. The function also includes the attribute to display on the x-axis “PVM” corresponding to the date, on the y-axis “KmYht” related to the total kilometres driven every day and the attribute “Auto” to assign a different colour to each taxi. The other functions added are: “geom_point” that assigns a dot to each value; “geom_ine” that links each dot of the same colour; “scale_color_manual” to define which colour we want to use; and “ggtitle” to assign a title to the graph and “labs” to define a name for the axis to help to interpret the scatter plot. The final commands shown below, repeated for the three different data frames concerning the three different months, provide the scatter plot shown in figures 23, 24 and 25:

```
qplot(data = sep, x = PVM, y = KmYht, color = Auto ) +
geom_point() + geom_line() + scale_color_manual(values=c("red",
"lightblue", "green", "sienna", "orange", "royalblue2", "pur-
ple", "yellow", "black")) + ggtitle("Total amount of kilometres
driven by each taxi in November") +
labs(x="Date",y="Kilometres")
```

On the resulting scatter plot, we can find new interesting information such as the kilometres that have been driven in a day. Since the graphs produced are related to a single month, it is also quite easy to determinate the exact value, the total amount of kilometres driven, of a specific taxi in a specific day. With this scatter plot we can see simultaneous the performances of the all nine cars. The coloured continuous line connecting the dots helps to understand the overall achievement of a taxi. For example, we can easy identify the purple line as one of the taxis that has been driving less, and by reading the legend we can see that the car “SH27” was the one that overall drove much less. If we would apply this kind of analysis to a much more specific data frame where we have the kilometres for each ride, we could easily spot the longest rides and perhaps detect if it is always the same customer going with a certain frequency to the same place.

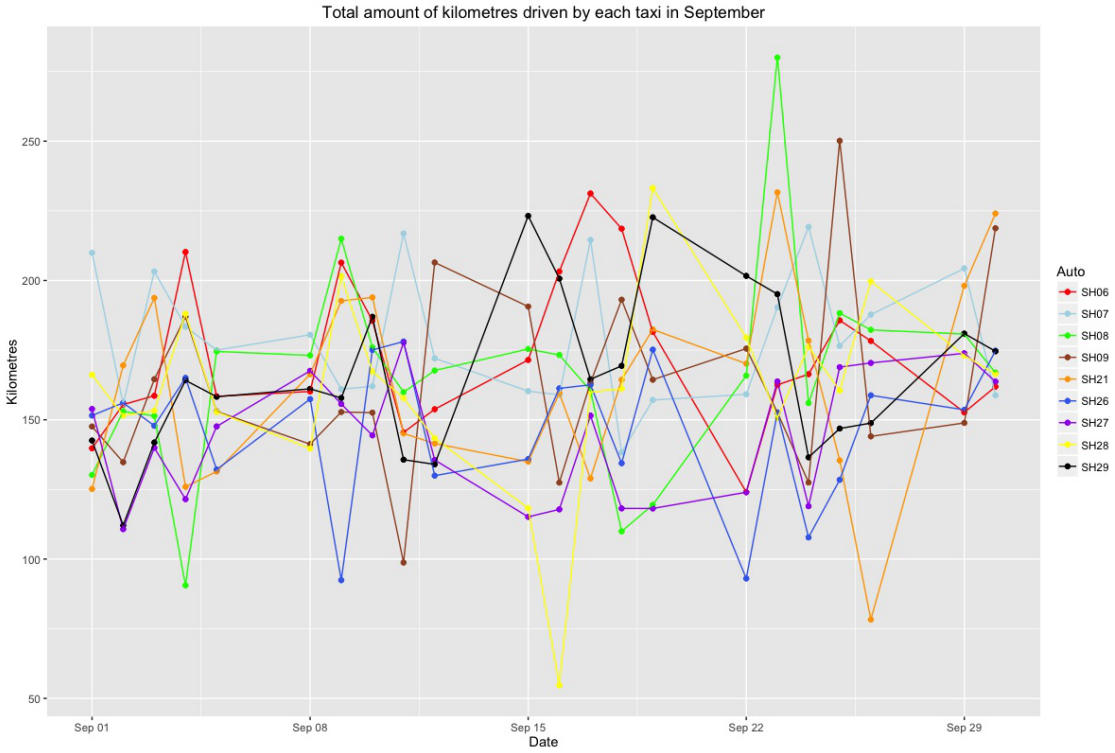

Figure 23. Scatter plot reproducing the kilometres driven by each taxi in September.

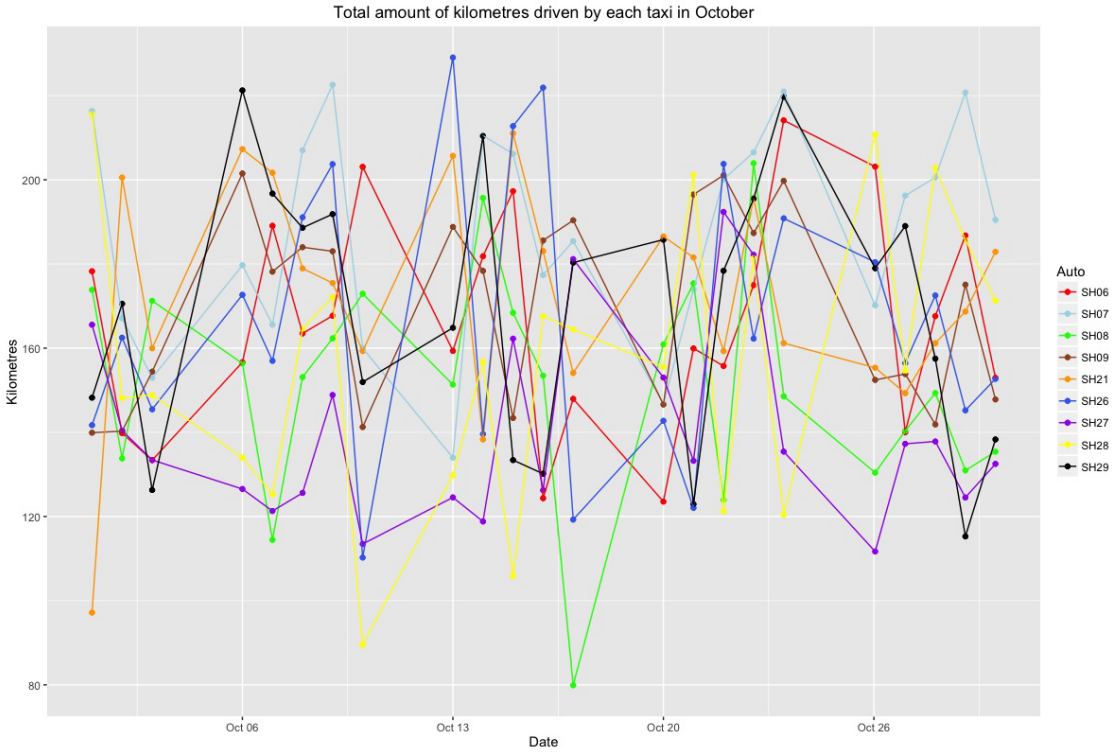

Figure 24. Scatter plot reproducing the kilometres driven by each taxi in October.

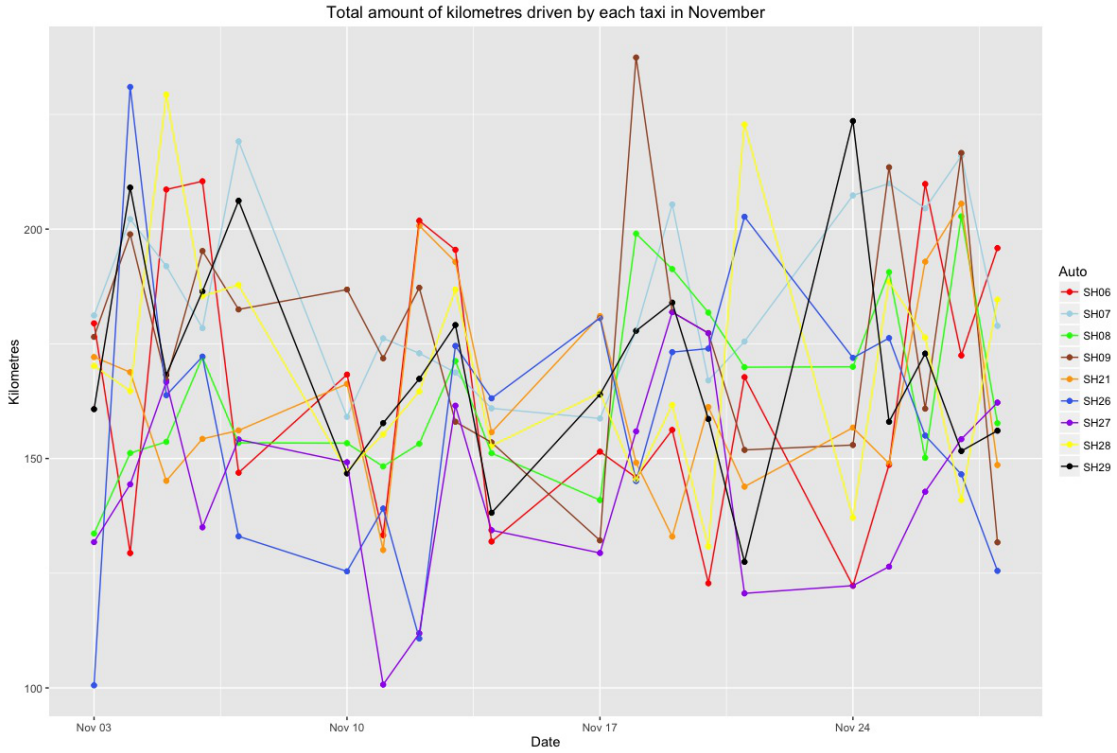

Figure 25. Scatter plot reproducing the kilometres driven by each taxi in November.

5 Evaluation of results

The aim of this section is to illustrate the results achieved during the development of the project. The meaning of this phase, called evaluation in the CRISP-DM model, is to present the results and also to evaluate if they match the goal.

The bar chart provided in figure 19 that gives the total kilometres driven by each taxi, appears very easy to read. However, in order to get a more accurate result, we should use an attribute that collects the corresponding monetary value. Since this attribute is not present in the database used it was decided to provide a more accurate analysis by using the total number of rides by each taxi; since for each ride there is a fixed fee attributed. The comparison is visible in figure 20. With this information, the manager of the company is able to find more accurate results by multiplying the value with a fixed fee. However, this attribute regarding the total kilometres, even if is directly related with the fee of the ride, does not take into consideration the waiting time; for example, the time that a car is waiting for a customer is an element that is very common with these kinds of rides where most of the customers have physical challenges. This element affects determining the total fee and consequentially the profitability of the business. For this reason, in order to know exactly the profitability it is necessary to investigate further by using an attribute that collects data regarding the fees. With this information the bar chart provides the manager of the company the ability to make more in depth conclusions by comparing the results with the drivers of the taxis, which is an attribute missing in the database used for these studies. He may be able to deduce that a driver is faster or slower than others.

As mentioned before, we were not able to predict where rides were coming from, because the required attribute collecting information about addresses presents inconsistent and corrupted data due to driving instructions or the name of the buildings being included as part of the address or due to a missing building number. In case of multiple rides, there are up to six addresses in the same field, this happens when three customers share the same taxi.

However, with these studies we were able to establish that in the three months selected from September until November 2014, the taxis that drove more kilometres had an afternoon shift as can be seen from figures 16 and 17, which respectively show the total amount of kilometres driven by the taxis and their shifts. This result was not really

expected, as we would not expect to get such a clear distinction between the morning and afternoon shifts.

Another point of this investigation concerned whether if on some days of the week there were more rides than others. We expected that Monday and Friday would be busier than other days. However, the results illustrated by figure 21 show that overall, in the three months selected, there is an equally amount of rides every day of the week. This result needs to be clarified, however, as there are not necessarily the same amount of customers every day of the week, as the taxi company has an agreement with the company organizing the rides that guaranteed a minimum number of rides. Probably, if on a day of the week there are more customers requiring a taxi, those rides are distributed to the taxis driving also normal rides. All the same, this result can be used if the rides assigned meet the agreement.

During the research phase one of the most important goals was to provide a graph that could help a manager of the taxi company to read the data more easily. For this reason we dedicated the last part of the modelling phase to implementing a different chart showing different points of view. One chart produced, shown in figure 22, is a plot box grouping the total kilometres driven every day by each taxi. This, easy to read, graph allows us to summarise the data and to have at a glance the overall situation of the period selected.

The others graphs give a more specific view, and for this reason it was considered necessary to split the data frame by month. The results, as shown in figures 23, 24 and 25, are three scatter plots reproducing the total kilometres driven every day by each taxi. With these scatter plots it is possible to compare all the taxis at the same time, to see the exact day inspected and to know the total kilometres driven.

6 Conclusion

Overall, it is possible to establish that the results achieved and graphs provided can help the taxi company to interpret the profitability of working with these special rides provided by the city. For privacy reasons the research was conducted by ignoring monetary values and driver ID numbers that could have delivered more accurate results. However, using other values, such as kilometres driven by the taxis, we were able to present an appropriate analysis that showed aspects from different perspectives; sometimes more detailed and sometimes more general.

In the period taken under inspection, it was considered as a parameter to establish the best taxi performance and the total kilometres driven by the taxi from the moment that the car is booked until arriving to the picking point and the kilometres driven with the customer. By analysing these values, a substantial difference between the car SH07 that was driving about 1,400 kilometres over the average and car SH27 that was driving approximately 1,400 less than average was detected. It should be kept in mind that in order to get the most accurate result it is necessary to use the amount of total fee for each ride as a parameter. This attribute is available on the database but it was omitted for this research for privacy reason.

For limitation of the database used for this research, it was not possible to predict if there are rides coming with regularity since the attribute concerning the address was defined corrupted. Somewhat surprisingly, other results provided show that there are no particular differences between days of the week. This result does not necessarily mean that every day of the week the same amount of taxis is needed but rather that the nine cars get overall a similar number of rides. However, by comparing number of total kilometres driven by the taxis and shifts, it was revealed that taxis driving afternoon shifts tend to drive more kilometres.

Finally, the results provided with this study and, in particular, the different methods applied to the sample of three months from September until November 2014, can be used during other periods of the year and they represent a tool to measure and to compare the performances of the nine cars. It could be interesting to investigate more and apply these methods frequently, like every month and see if similar results are achieved and eventually establish, if differences are generated by the skills of the drivers or work shifts or other factors.

Overall RStudio environment, including packages, represents a valid powerful and modern instrument for statistical computing and data analyse. It is ideal for taking under inspection information collected in databases. The usage of this platform is also suitable for small business with the open source version. RStudio allows to create interactive and clear graphs deploying results of the data mining process which are easy to read and handy to use for reports.

References

1 Borgman C. Big Data, Little Data, No Data – Scholarship in the Networked World. Cambridge, Massachusetts, United States: The MIT Press. 2015

2 Anderson. C. The long tail.
URL:http://www.longtail.com/about.html.
Accessed: 30 March 2016.

3 Fox P, Harris R. ICSU and the challenges of data and information management for international science. Data Science Journal 2013;12:10-11.

4 WhatIs [online].
URL: http://whatis.techtarget.com/definition/metadata.
Accessed: 30 March 2016.

5 Guenther R, Radebaugh J. Understanding metadata [online]. Bethesda, Montgomery, United States: NISO Press. 2004.

6 RStudio [online].
URL: https://www.rstudio.com/.
Accessed: 30 March 2016.

7 The R Project for Statistical Computing [online].
URL: https://www.r-project.org/.
Accessed: 30 March 2016.

8 GNU Operating System [online].
URL: http://www.gnu.org/.
Accessed:30 March 2016.

9 R-bloggers [online].
URL: http://www.r-bloggers.com.
Accessed: 30 March 2016.

10 The Comprehensive R Archive Network [online].
URL: https://cran.r-project.org/.
Accessed: 30 March 2016.

11 Gartner glossary[online].
URL: http://www.gartner.com/it-glossary/.
Accessed: 30 March 2016.

12 Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discov ery in Databases. AI Magazine. 1996;17:39-41.

13 Azevedo A, Santos MF. KDD, SEMMA and CRISP-DM: A parallel overview. Ams terdam, The Netherlands. IADIS European Conference on Data Mining. 2008[online].
URL:
https://www.researchgate.net/publication/220969845_KDD_semma_and_CRISP-DM_A_parallel_overview.
Accessed: 30 March 2016.

14 Brown I. Data Exploration and Visualisation in SAS Enterprise Miner [online]. United Kingdom. 2015.
URL:
http://www.sas.com/content/dam/SAS/en_gb/doc/other1/events/sasforum/slides/day2/I.%20Brown%20Data%20Exploration%20and%20Visualisation%20in%20SAS%20EM_IB.pdf.
Accessed 30 March 2016.

15 Larose DT. Discovering knowledge in data: an introduction to data mining. Wiley-In terscience. 2004.

16 Data Integration Info [online].
URL: http://www.dataintegration.info/etl.
Accessed: 30 March 2016.

17 McCallum QE. Bad data handbook Mapping the World of Data Problems – Q. Ethan McCallum. O'Reilly Media. 2012.

18 Abhinav A. Programming in R [online].
URL:
http://rstudio-pubstas-tic.s3.amazonaws.com/21347_418bc228038d4e94815018ad415bba49.html.
Accessed: 30 March 2016.

19 Stat Trek – Teach Yourself statistics [online].
URL: http://stattrek.com/statistics/charts/boxplot.aspx?Tutorial=AP.
Accessed: 30 March 2016.

20 Plotly [online].
URL: https://plot.ly/.
Accessed: 30 March 2016.

21 Quick-R [online].
URL: http://www.statmethods.net/advgraphs/ggplot2.html.
Accessed: 30 March 2016.